5th May 2020

Credit Card Fraud Detection

Anushka Patil[1], Sreya Venkatesh[2], Dr. Angel Arul Jothi[3]
[1,2] Student, Dept. of Computer Science, Birla Institute of Technology and Science, Pilani, Dubai Campus
[3] Instructor In Charge, Dept. of Computer Science, Birla Institute of Technology and Science, Pilani, Dubai Campus

Abstract: *The research is focused on the real-world credit card fraud detection. As the number of credit card transactions increase in the modern day, so does the fraudulent activities with different motives. With the rise of digitalization, the biggest challenge faced by the businesses is to carry out the transactions even when the presence of the buyer is not possible. This makes it next to impossible to always verify and carry out the transaction, thus increasing the number of fraudulent activities. Through this study we aim to build a model than can classify a transaction as fraudulent or legitimate. Various experiments were conducted using three classifiers, namely; Naïve Bayes, Logistic Regression and K-Nearest Neighbor. Evaluation metrics such as precision, accuracy, recall and f1 score are used to indicate the best possible model.*

1. INTRODUCTION

1.1 GENERAL INTRODUCTION TO THE PROJECT AREA

Credit card fraud is defined as the illegal usage of another's credit card information to obtain goods, services or anything of value. Using the credit card number without actually possessing the credit card is also a form of credit card fraud. Global Payments Report states that the credit card is the most popular payment method worldwide compared to other such as bank transfer and e-wallet. As this industry continues to grow and offer credit to more people, fraud will also be more prevalent.

At present, data mining is the most prevalent method to tackle fraud due to its effectiveness. The various methods used in fraud detection include, the Hidden Markov model, Bayesian algorithm, K-Nearest neighbor, Fuzzy logic, Ensemble classification, Support Vector machine etc. the two main categories that statistical fraud detection is divided into are: Supervised and Unsupervised learning. In supervised fraud detection, samples of legitimate and fraudulent transactions are estimated so as to be able to classify whether the recent new transactions are legitimate or not. Whereas in unsupervised fraud detection, probable cases of fraudulent transactions are identified with the help of outliers or illegitimate transactions. Both of these methods only help in predicting the probability of a fraudulent transaction. Banks use data mining to detect patterns in a group and extract unknown relationship in the data.

1.2 OBJECTIVES

The objective of this project is to analyze and study the real world data. We aim at understading all the challenges and incorporating their solutions so as to achieve utmost accuracy using the naïve bias model in data mining. Through the process of feature extraction which is executed carefully, we target to attain best possible set of attributes to avoid overfitting and provide best results.

1.3 MOTIVATIONS/CHALLENGES

These days customers most preferred mode of payment is the credit card for convenience in online shopping etc. Due to this past couple of years has seen an alarming trend of credit card breaches being reported. With the development of numerous credit card fraud detection methods perpetrators of fraud are also advancing their methods to avoid being detected. Thus, it is highly critical to develop better credit card fraud detection methods to combat the increasingly intelligent fraudulent activities. Globally about $24.26 Billion was lost due to payment card fraud in 2018. There is a credit card fraud increase of 18.4 percent worldwide in 2018 and this is still currently on the rise. The number one identity theft fraud is credit card fraud. Fraud is a national and global problem and with such statistics it is only vital that the world cope and implement measures to combat these fraudsters.

Credit card fraud detection is a vastly explored area of fraud detection and to detect fraudulent behavior it relies on analysis of recorded transactions automatically. More importantly credit card fraud detection is plagued with various problems and constraints. Given below are some of the challenged faced while detecting credit card fraud:

i) Unavailability of real data set: This is one of the biggest issues to be dealt with as banks and financial institutions do not give away their customers transaction details readily due to concerns over privacy.

ii) Determining the appropriate evaluation parameters: False positives and False negatives are two common and basic measures to detect fraud. Both of these methods have an opposite relationship; while one decreases the other increases. Accuracy is not appropriate to detect fraud as the dataset is highly imbalanced. Hence, all the fraudulent transactions can be misclassified due to high accuracy.

iii) Size of the dataset: There are more than a million credit card transactions processed every day. Highly efficient techniques that scale well and a lot of computing power is necessary to such enormous amounts of transactions.

iv) Unbalanced dataset: Credit card fraud datasets are extremely skewed with legal and fraudulent transactions varying several times.

v) Dynamic behavior of fraudster: It cannot be expected that the fraudster will always behave in the same way to get through the fraud detection systems. Fraudsters change their behavior and modify their fraud style to get through the increasingly complex and developed detection systems.

But even with these concerns and challenges credit card fraud is widely studied.

1.4 ORGANIZATION OF THE REPORT

The report is broadly divided into 3 parts. The first part consists of all the past research and study related to the current topic of our project. It is summarized in the form of a literature survey for a brief overview. The second part of the report includes data pre-processing, feature extraction and the understanding and learning of the different models.The third part of the report implements three models, Naïve Bayes, Logistic Regression and K-Nearest Neighbor and uses different evaluation metrics to conclude the best possible model for fraud detection.

2. RELATED WORK

Many researchers have conducted studies concerning application of data mining techniques in detecting credit card frauds. They extend the use of different methods and algorithms and as to which algorithm best suits the problem is still an ongoing debate.

Several studies throughout the years applied neural networks to detect credit card frauds. Ogwueleka[1] in 2011 used artificial neural networks combined with a rule based component. The developed Credit Card Fraud Watch, which would run along with the banking software to detect anomalies used four clusters; low, high, risky and high risk, instead of the two-stage model commonly used in other studies. This reduced the chances of legitimate transactions being incorrectly labeled as fraudulent and in turn improved accuracy.

To obtain better solutions, in [7] Genetic Algorithms are used. Here, fraudulent transactions are created with the help of given sample dataset. Using this method, fraud can be detected soon after the credit card transaction occurs instead of waiting for days. The experiment setup consisted of four steps; data input,computing critical values,generating critical values after some generations and then generating fraud transactions using this algorithm.

A decision tree Hunt's algorithm was adopted by Dr R.DHANAPAL and, GAYATHIRI.P [6] to detect the fraud and trace email and IP address. Various splitting criteria such as Entropy,Information Gain and Gini are studied and an Information gain based Decision tree is built.

Because of Globalization and emergence of the internet, online transactions have increased rapidly. [3] uses Fuzzy logic and K-means Algorithm to deal with these online credit card frauds. K-means, which is an unsupervised algorithm, was applied to the dataset using matlab and the fuzzy function was used to identify behavioural change of the customer. The attributes chosen to be analysed were transaction number, amount, shipping address and customer address.

Masoumeh and Pourya [4] investigated four techniques of fraud detection; Naive Bayes, Support Vector Machines, KNN and Bagging ensemble classifier.This paper uses Bagging ensemble classifier based on decision trees. They used the real world data set obtained from UCSD-FICO competition that contained a ratio of legitimate and fraudulent transactions as 100:3. Due to highly imbalanced dataset, the dataset was divided into four parts. The metrics like error rate and accuracy are not used as they might be highly biased due to imbalance in

the dataset. Instead False Alarm Rate, Balanced Classification Rate, Fraud Catching Rate and Matthews Correlation Coefficient are used.

The authors of [10] used unsupervised method for credit card fraud detection. The dataset contains the amount spent and the location of the transaction made by the customer. It performs cluster analysis and uses Euclidean distance function.

Nurul Malim, Saravanan and Ong Shu Yee [8] used an assortment of machine learning and data mining methods. To identify the patterns of the normalized data, data mining was required. Then they were correctly labelled as suspicious and non-suspicious. Then, machine learning (ML) techniques using classifiers were implemented to analyse and forecast the transactions as suspicious and non-suspicious automatically. This paper further studies J48 classifiers, Tree Augmented Naive Bayes (TAN), logistics and Bayesian network(K2).

[5] and [9] uses Random Forest based Classifiers. Both of the journal papers provides the relevant information on what Python libraries are to be used namely; Numpy, Pandas, Matplotlib and Sklearn.For feature extraction [5], the researchers used SMOTE(Synthetic Minority Oversampling Technique), to increase the number of instances in the dataset in a balanced manner. [9] claimed that Random Forest would work better with large training set but the speed to test data might suffer.

Although numerous researches have been conducted, there are many challenges to be addressed to attain best performance. Randula [2] discusses issues such as the imbalanced nature of data, missing data, overlapping data and new emerging fraud patterns. It provides an analysis of various algorithms used in credit card fraud detections and states its problems, limitations and future scope.

3. METHODOLOGY

3.1 BLOCK DIAGRAM / ARCHITECTURE

The dataset consisting of credit card transactions are obtained from the source (in this paper, from Kaggle). First, data set cleaning and feature extraction is performed. This involves, looking for missing values, outlier removal, converting variables to usable forms etc. Then, the dataset is divided into two parts, training set and test set. The model is built using the training set and predictions are made using the test set on the built model. Then using the evaluation metrics, the best scheme to build the model is chosen.

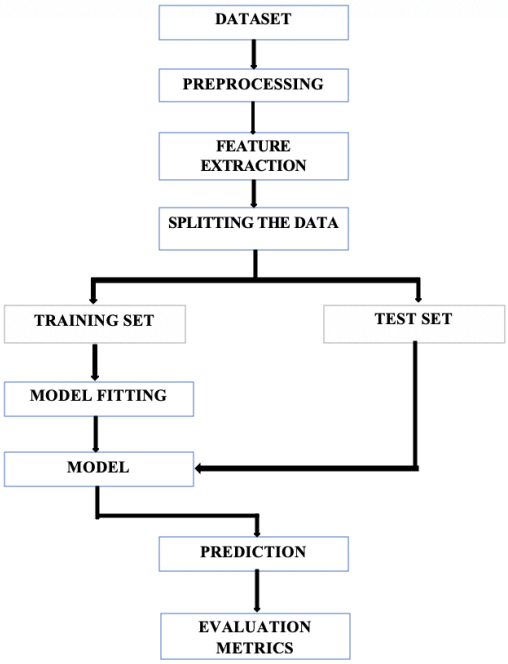

Fig. 1: Block Diagram

3.2 MODULE DESCRIPTION

In this section, we will explain each step of the block diagram above along with the input and output of each step.

3.2.1 PREPROCESSING

Input: Raw dataset
Output: Normalized data
Preprocessing is needed to transform raw data obtained from the real world to an understandable format. In this research a series of steps were conducted that includes; checking for missing values and replacing or removal of them, outlier detection, duplicates removal, class identification and normalizing the data.

3.2.2 FEATURE EXTRACTION

Input: Normalized data
Output: Feature Vector
Feature extraction is a method of reducing the dimensionality of the data without losing important attribute dependencies on the class label. Some attributes show similar distributions for both the classes: Legitimate and Fraud. We can eliminate these attributes to reduce complexity and avoid overfitting.

3.2.3 SPLITTING THE DATA

Input: Feature Vector
Output: Training set and Test set

This study includes 284,807 transactions, we would split the data into test set and training set. As the data is highly imbalanced we would try to split it using train_test_split() function to distribute data reliably and avoid underfitting.

3.2.4 MODEL FITTING

Input: Training Set
Output: Model
The model is prepared by learning the training set. The classifiers that are chosen for this study are Naïve Bayes, KNN and Logistic Regression.

3.2.5 MODEL EVALUATION

Input: Test Set
Output: Predicted class values for test set and statistics
This is one of the most important part of the model building. After a series of experiments the model that is the most suitable for future work is chosen to be the ideal model. To evaluate model performance, we use the data not seen by the model, that is the test set to avoid overfitting. The chosen metrics to analyze the model performance in this study are Recall, Precision, Accuracy and F1 score.

4. DATASET DESCRIPTION

In this research experiment, we would be using real life data obtained from Kaggle. The dataset obtained, contained 284,807 transactions made by European credit cardholders over a period of two days in September 2013.
The dataset obtained, contained 492 fraud transactions which accounts for only 0.172% of all transactions. This implies that the data set is highly imbalanced.

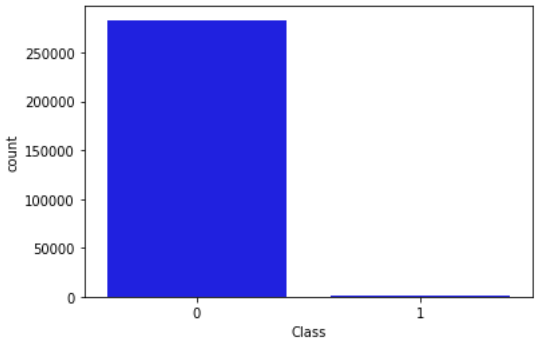

Fig. 2: Class distribution in the dataset

This dataset has 31 attributes:
i) Time – This tells the elapsed number of seconds between current transaction and the first transaction in the dataset
ii) V1-V28 – These attributes are a result of PCA Dimensionality reduction. This was done to prevent the client sensitive information.
iii) Amount – amount being transmitted
iv) Class – the class label is 1 if the transaction is fraudulent and 0 if its legitimate.

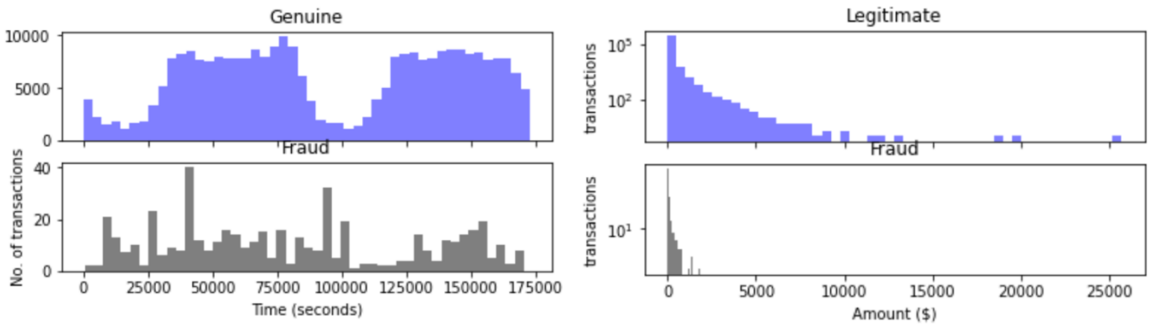

Fig. 3: Time and Amount feature distribution over fraud and legitimate classes

The dataset contains no null values, but it does contain duplicate transactions. 1081 duplicate rows are removed during preprocessing.
A Correlation matrix is formed and is then noticed that all features are related to the class variable that is our target variable. Hence, initially no features are dropped.

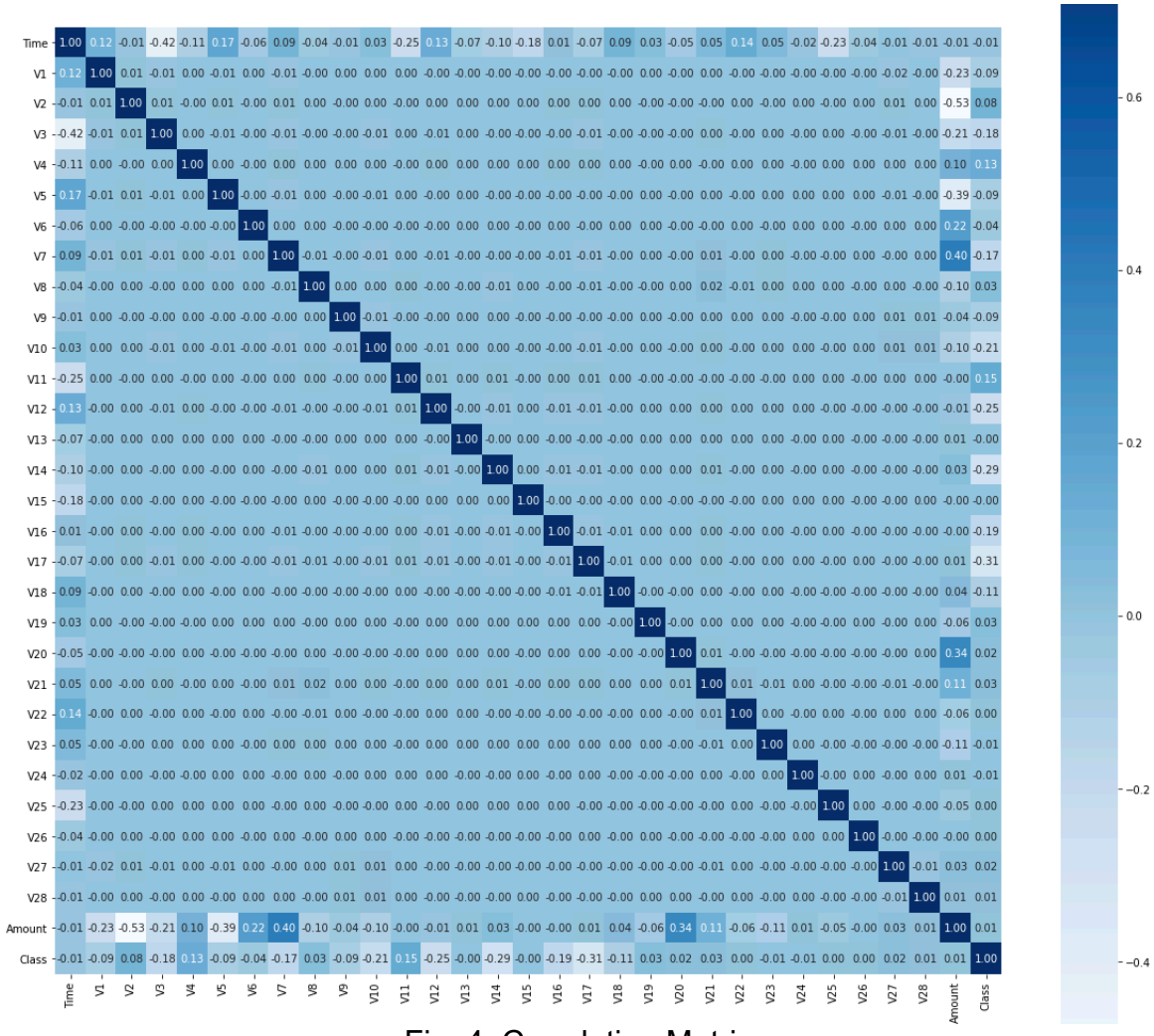

Fig. 4: Correlation Matrix

Next, and the most important step, the features are standardized into normal distribution using the StandardScaler. The features now have mean ‘0’ and standard deviation of ‘1’. We do so to make sure no one feature dominates the other features due to the presence of large magnitude value.

5. EVALUATION METRICS

In this study, Accuracy may not be a good evaluation metric due to highly imbalanced data set. The number of fraudulent transactions is far less in number when compared to the. Legitimate transactions. Therefore, F1 Score is a more reliable metric.

Confusion matrix

It is a table that describes the classifiers performance on a test data collection where the real values are known.

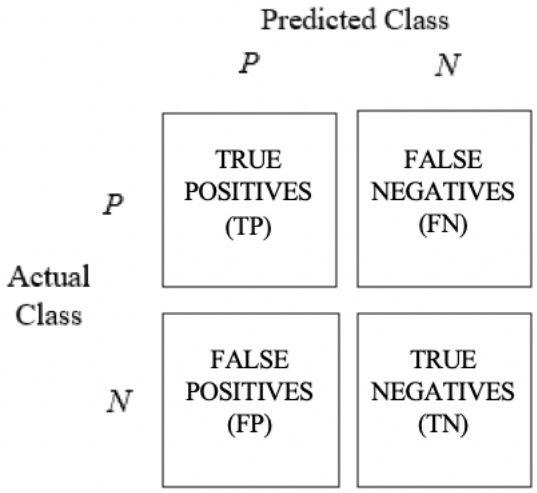

Fig. 5: Confusion Matrix

Table 1: Evaluation Metrics

Metric	Formula	Description
Recall	$= \frac{TP}{TP + FN}$	Recall is defined as the ratio between positive observations that are correctly predicted and all the actual positive observations.
Precision	$= \frac{TP}{TP + FP}$	Precision is the ratio between positive observations that are correctly predicted and all the predicted positive observations.
Accuracy	$= \frac{TP + TN}{TP + FP + FN + TN}$	Accuracy is defined as the observations that are correctly predicted upon total number of observations.
F1 Score	$= \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$	It is basically a statistical measure to rate performance and is defined as the harmonic mean between recall and precision. It measures how accurate a test is.

6. EXPERIMENTS CONDUCTED

In this research, three classifiers were used, namely; Naïve Bayes, Logistic Regression and K-Nearest Neighbor. Further, several iterations were performed for each of them with slight changes in the training features and parameters.

6.1 NAÏVE BAYES

Naïve Bayes is a commonly used method for detecting credit card fraud. There are two probabilities that are stored by a learned Naïve Bayes model; probability of each class in training dataset and (conditional) probability for each input given each class value.

Formula: - P(h|d) = [P(d|h) * P(h)]/ P(d)

P(h|d) is "posterior probability" = prob. of hypothesis h for given data d, P(d|h) is the prob. of data d given hypothesis h is true, P(h) is "prior prob of h" being true regardless of data and P(d) is prob. of data regardless of hypo

This algorithm calculates posterior probability for all possible hypothesis and selects the one with maximum probability. This maximum probable hypothesis is called MAP (maximum a posteriori): MAP(h) = max(P(h|d).

Experiment 1: Naïve Bayes with all the features

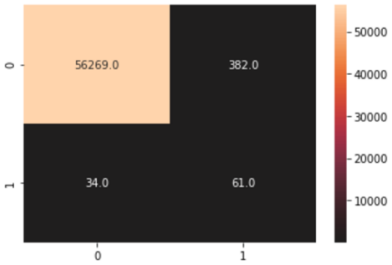

Fig. 6: Confusion Matrix for Experiment 1

Experiment 2: Naïve Bayes with features that have less correlation with the class label dropped ('V22','V26','V25','V15','V13','V23','V24')

Experiment 3: Naïve Bayes with features ‘Time’ and ‘Amount’ dropped in addition to the features dropped in experiment 2

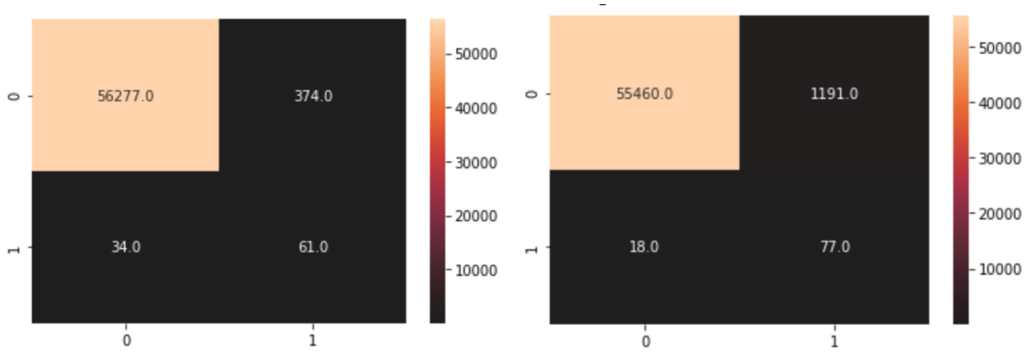

Fig. 7,8: Confusion Matrix for Experiment 2 and 3 respectively

6.2 LOGISTIC REGRESSION

Before the Logistic Regression is performed, the highly imbalanced dataset is first under sampled so that it contains equal number of fraudulent and legitimate transactions in the training set. This helps in avoiding overfitting by providing a balanced dataset. Logistic Regression is a statistical method which is well known for the prediction of binomial or multinomial outcomes. In the context of this research, Binomial Logistic Regression is used which is limited to the models that have binary class labels as the target field. Here, the target variable will either be corresponding to a fraudulent or legitimate transaction. Before the Logistic Regression is performed, the highly imbalanced dataset is first under sampled so that it contains equal number of fraudulent and legitimate transactions in the training set. This helps in avoiding overfitting by providing a balanced dataset. C is the inverse regularization parameter. This adds the penalty corresponding to the increase in magnitude of the values of the features. This is done so as to reduce the overfitting

Experiment 4: Logistic Regression with C=0.1 and using the entire dataset
Experiment 5: Logistic Regression with under sampling and C=0.1

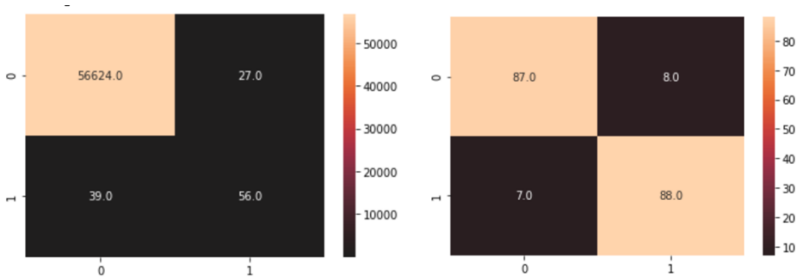

Fig. 9,10: Confusion Matrix for Experiment 4 and 5 respectively

6.3 K-NEAREST NEIGHBOUR

KNN is one of the simplest and commonly used methods for classification based on feature similarities. It is a lazy learner, in opposed to the eager learners. It does not create generalizations based on the learnt training data, in turn makes the learning phase very quick. Classification of an object is decided by the majority votes of its ‘k’ nearest neighbors and hence takes a class which is the most common in its neighborhood. Although it is an easy and quick process, it is highly influenced by the irrelevant features.

Experiment 6: KNN with k=’100’
Experiment 7: KNN with k=’5’

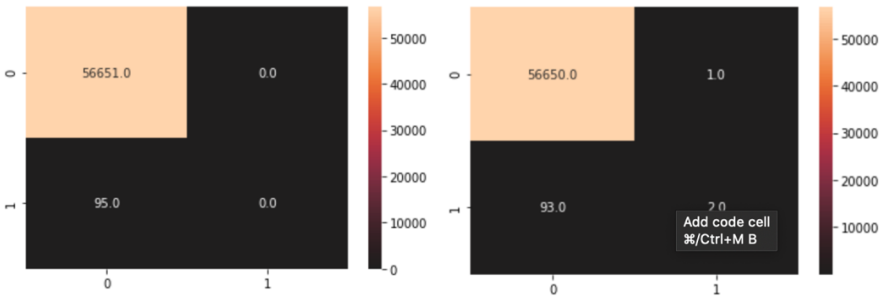

Fig. 11,12: Confusion Matrix for Experiment 6 and 7 respectively

In all of these above 7 experiments, the processed data is used. The dataset is divided into training set (80%) and test set (20%) and then the classifier model is built using the training set. From Scikit-learn machine learning library in python, we import useful classifiers such as GaussianNB, LogisticRegression and KneighborsClassifier to build the models for Naïve Bayes, Logistic Regression and KNN respectively.

7. RESULTS AND DISCUSSIONS

This study uses 7 experiments to determine the best possible model for predicting credit card frauds. All of these have achieved accuracy greater than 90% but accuracy is not a reliable measure for this study because of the highly imbalanced dataset with very few fraudulent transactions accounting to 0.17%. Due to this, Nearest neighbor is not a good classifier in this case. Even though it gives a high accuracy of 99.83%, it does not recognize the fraudulent class cases correctly as seen in confusion matrix, recall, precision and f1 score values. Naïve Bayes experiments imply that as irrelevant features are eliminated, the F1 score increases slightly but when we eliminate further it starts reducing again, implying underfitting. Logistic Regression is the most suitable model for this study. The model performance is improved additionally in experiment 5 by performing under sampling and taking equal proportions of fraudulent and legitimate transactions to train the model.

Table 2: Results of different experiments

Exp No.	Experiment description	recall	precision	F1 score	accuracy
1	Naïve Bayes (all features)	64.21%	13.76%	22.67%	99.26%
2	Naïve Bayes (eliminated less correlated features)	64.21%	14.02%	23.01%	99.28%
3	Naïve Bayes (eliminates Time and Amount)	81.05%	6.07%	11.29%	97.86%
4	Logistic Regression	58.94%	67.46%	62.92%	99.88%
5	Logistic Regression with under sampling	92.63%	91.66%	92.14%	92.10%
6	100-Nearest Neighbor	0%	0%	0%	99.83%
7	5-Nearest Neighbor	2.10%	66.66%	4.08%	99.83%

8. CONCLUSION AND FUTURE WORK

Overall, Logistic Regression is the best performing algorithm for our study with 92.14% f1 score and 99.88% accuracy.
Advantages of Logistic Regression:
i) It is a highly efficient algorithm that does not use many computational resources.
ii) It does not require scaling of the input training data.
ii) It is easy to implement and interpret.
Disadvantages of Logistic Regression:
i) It is not possible to separate data that is not linearly separable.
ii) It is not one of the most powerful algorithms and hence can be outperformed by more complex algorithms.
iii) This algorithm is vulnerable to overfitting in high dimensional dataset.

As a future work, instead of just using accuracy, recall, precision and f1 score as performance measures, more metrics would be considered and included. Also, more refined preprocessing and removal of outliers would be performed. Apart from this, the study can be extended to more powerful algorithms of neural networks.

REFERENCES

[1] Ogwueleka, Francisca. (2011). Data mining application in credit card fraud detection system. Journal of Engineering Science and Technology. 6. 311-322.

[2] Koralage, Randula. (2019). Data Mining Techniques for Credit Card Fraud Detection.

[3] Murugesan, Balamurugan & Mathiazhagan, P. (2015). Credit Card Transaction Fraud Detection System Using Fuzzy Logic and K-Means Algorithm. International Journal of Innovative Research in Technology 2349-6002. 2.

[4] Zareapoor, Masoumeh & Shamsolmoali, Pourya. (2015). Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier. Procedia Computer Science. 48. 679-686.

[5] Mohankumar, Karuppasamy, K. (2019). Credit Card Fraud Detection using Random Forest Technique. International Journal of Innovative Research in Science, Engineering and Technology 2319-8753. 8.

[6] Dhanapal, Gayathiri. (2012). Credit Card Fraud Detection Using Decision Tree for Tracing Email and Ip. International Journal of Computer Science Issues 1694-0814. 9.

[7] Ramakalyani K., Umadevi D. (2012). Fraud Detection of Credit Card Payment System by Genetic Algorithm. International Journal of Scientific & Engineering Research 2229-5518. 3.

[8] Sagadevan, Saravanan & Malim, Nurul & Yee, Ong. (2018). Credit Card Fraud Detection Using Machine Learning as Data Mining Technique. Journal of Telecommunication, Electronic and Computer Engineering 2289-8131. 4.

[9] Devi Meenakshi, Janani B., Gayathri S., Indira N. (2019). Credit Card Fraud Detection using Random Forest. International Research Journal of Engineering and Technology 2395-0056. 6.

[10] Agaskar, V. Megha Babariya, Shruthi Chandran, Namrata Giri. (2017). Unsupervised Learning for Credit Card fraud detection. International Research Journal of Engineering and Technology 2395 -0056. 4.